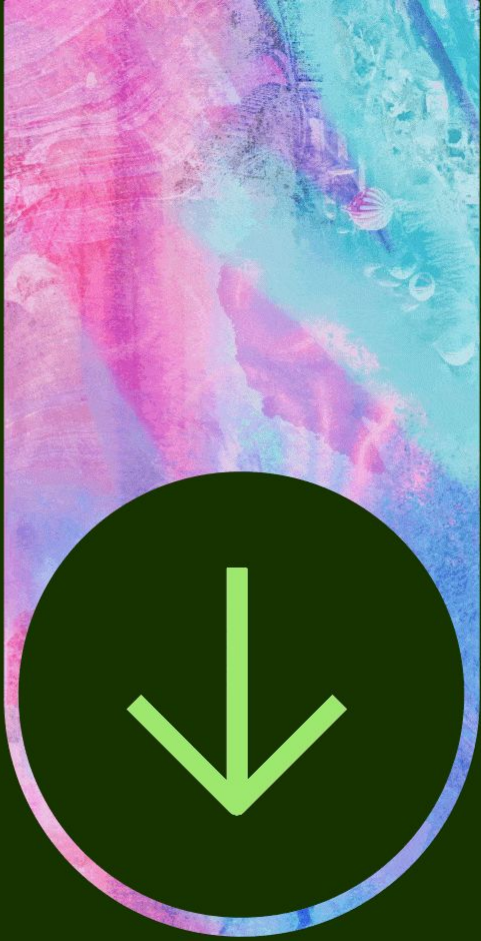


Graphs on Big Data

Why don't you use Neo4j?





Otto von Sperling

Data Scientist
Servicing Platform

Servicing customers

TALK

KNOW

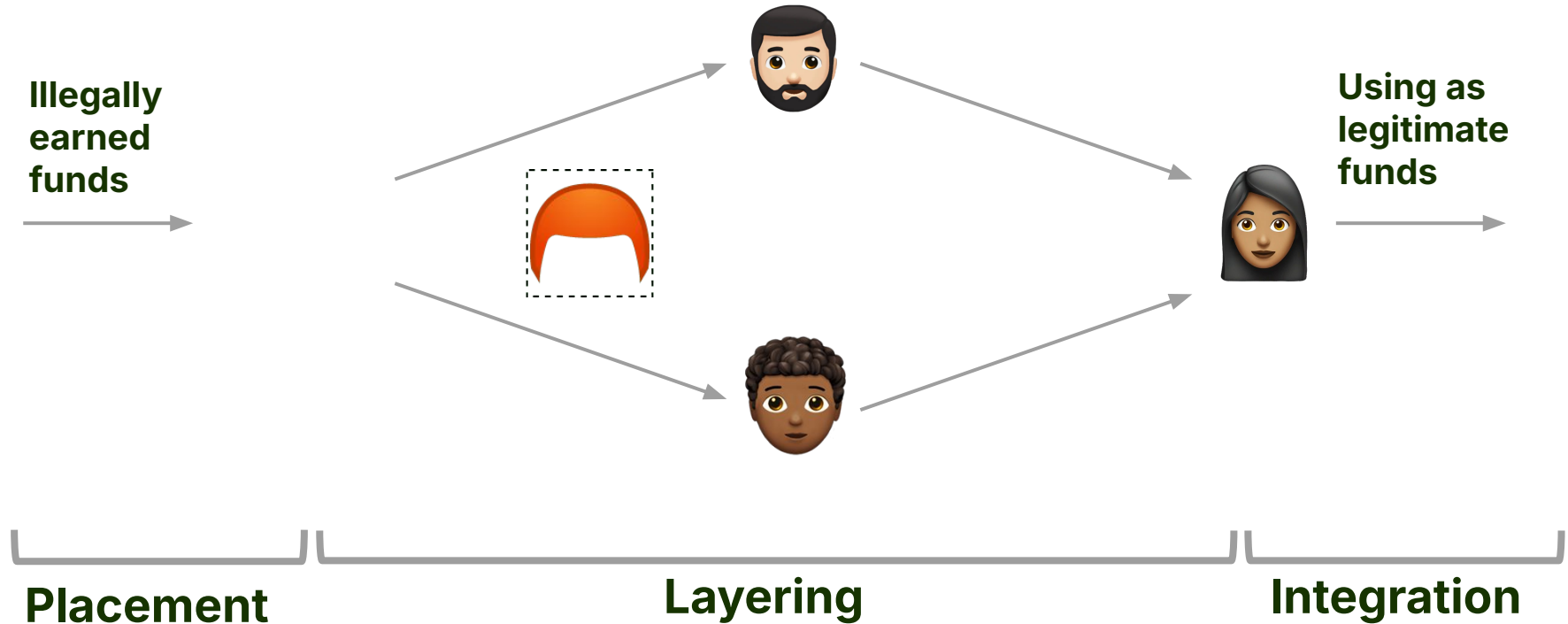
PROTECT



PROBLEM

Finding accounts linked to criminals

Money Laundering - simple example



Networks without graphs

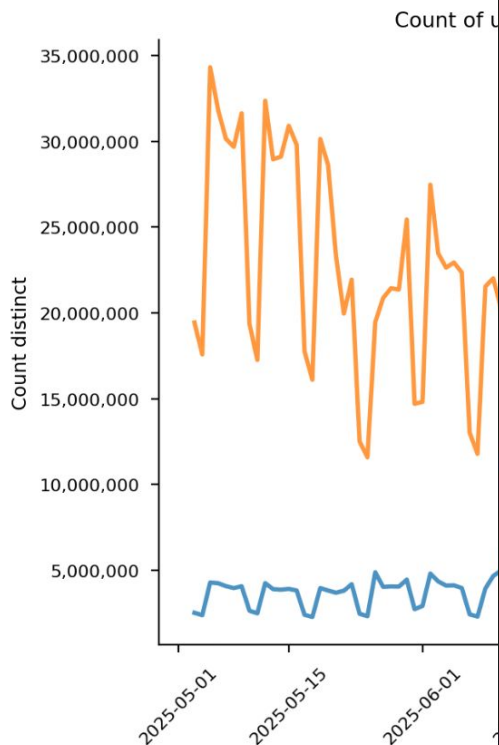
Feature Engineering

E.g. Ratio of identities marked as fraudulent in one's recipient list

Organic Art?

That slow down investigations

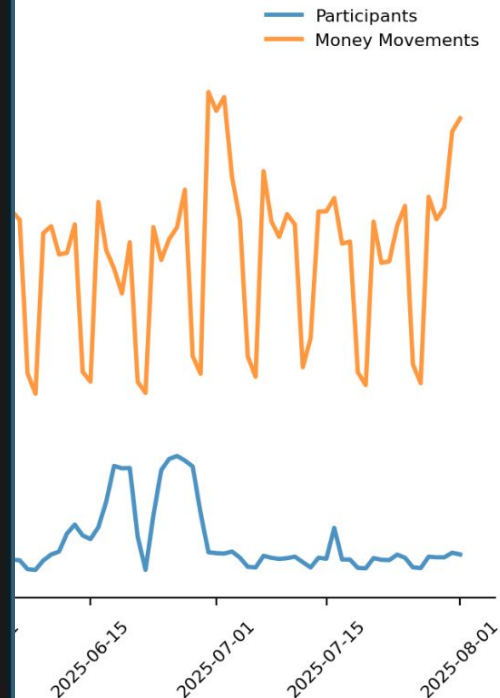
A lot of data



Exchange

shuffle records written: 1,944,023,621
shuffle write time total (min, med, max)
10.6 m (38 ms, 122 ms, 836 ms)
shuffle close time total (min, med, max)
2.41 h (640 ms, 1.9 s, 33.1 s)
shuffle sort time total (min, med, max)
25.0 s (1 ms, 5 ms, 94 ms)
records read: 1,944,023,621
local bytes read total (min, med, max)
5.4 GiB (473.2 KiB, 827.6 KiB, 1240.6 KiB)
fetch wait time total (min, med, max)
44.3 m (0 ms, 0 ms, 6.6 s)
remote bytes read total (min, med, max)
1027.4 GiB (151.0 MiB, 154.7 MiB, 159.0 MiB)
local blocks read: 168,060
remote blocks read: 32,106,384
data size total (min, med, max)
5.6 TiB (408.8 MiB, 1464.1 MiB, 1972.2 MiB)
number of partitions: 6,800
remote reqs duration total (min, med, max)
6.27 h (647 ms, 1.2 s, 33.7 s)
shuffle bytes written total (min, med, max)
1032.8 GiB (77.8 MiB, 259.1 MiB, 339.9 MiB)
shuffle merge time total (min, med, max)
3.8 m (0 ms, 0 ms, 1.0 m)

Volume of raw data

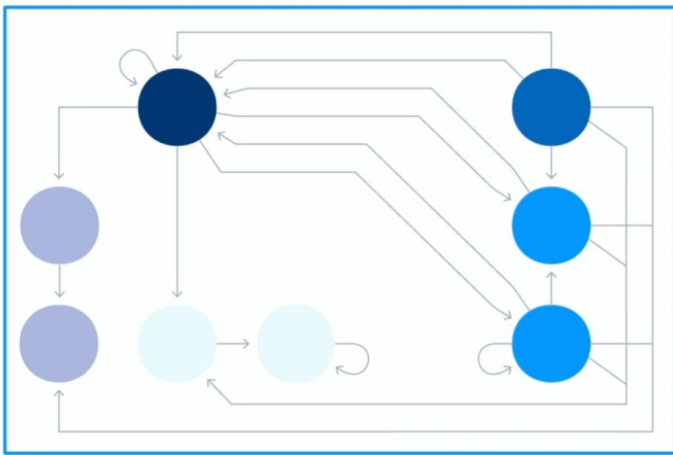


Neo4j

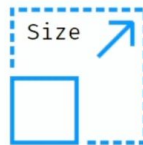
On big data?

Approach #2: Graph Native Sharding

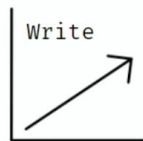
Invest in
Data Modeling



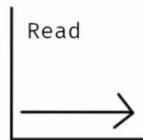
Get These Magical
Properties:



Unbounded size



Linear write throughput



Constant read performance

Hassle-free performance?



Vendors

Buy → Build → Bye-Bye

**ONE STEP
BACK**

What are databases for?

Consistency

Availability

Partitions

Easy to get a new DB?

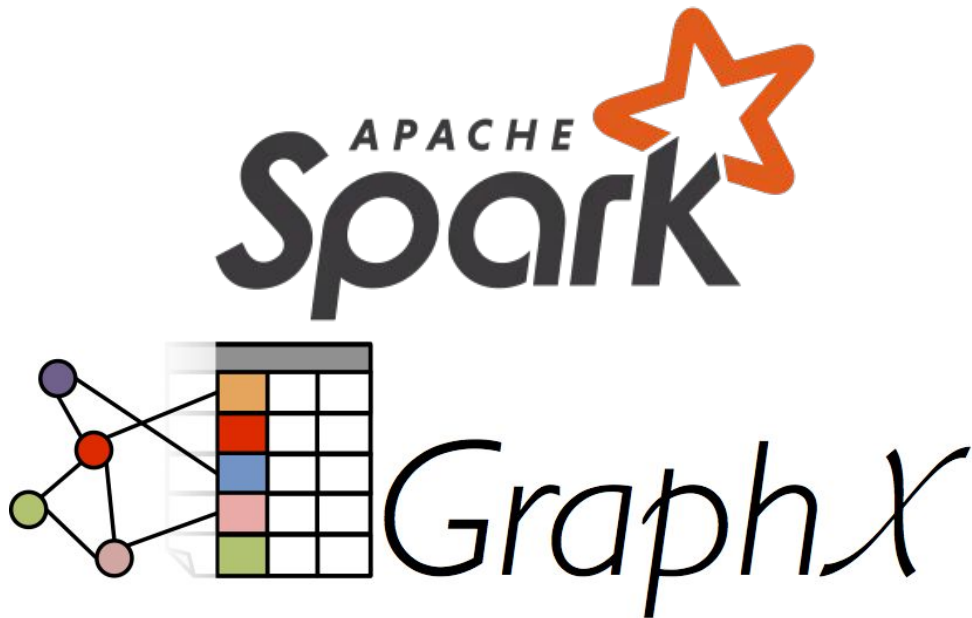


**ANOTHER
STEP BACK**

Alert investigation SLA?

one day, up to a week

Big data + longer SLAs...



Long live...

[DISCUSS] Deprecate

September 29, 2024

[ASF] Board Report: GraphX Deprecated

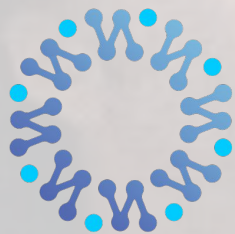
November 20, 2024

and new maintainers... ?

[3]: <https://lists.apache.org/thread/qrvo6xrt8zvp5ss73z5spt9q89r0htwo>

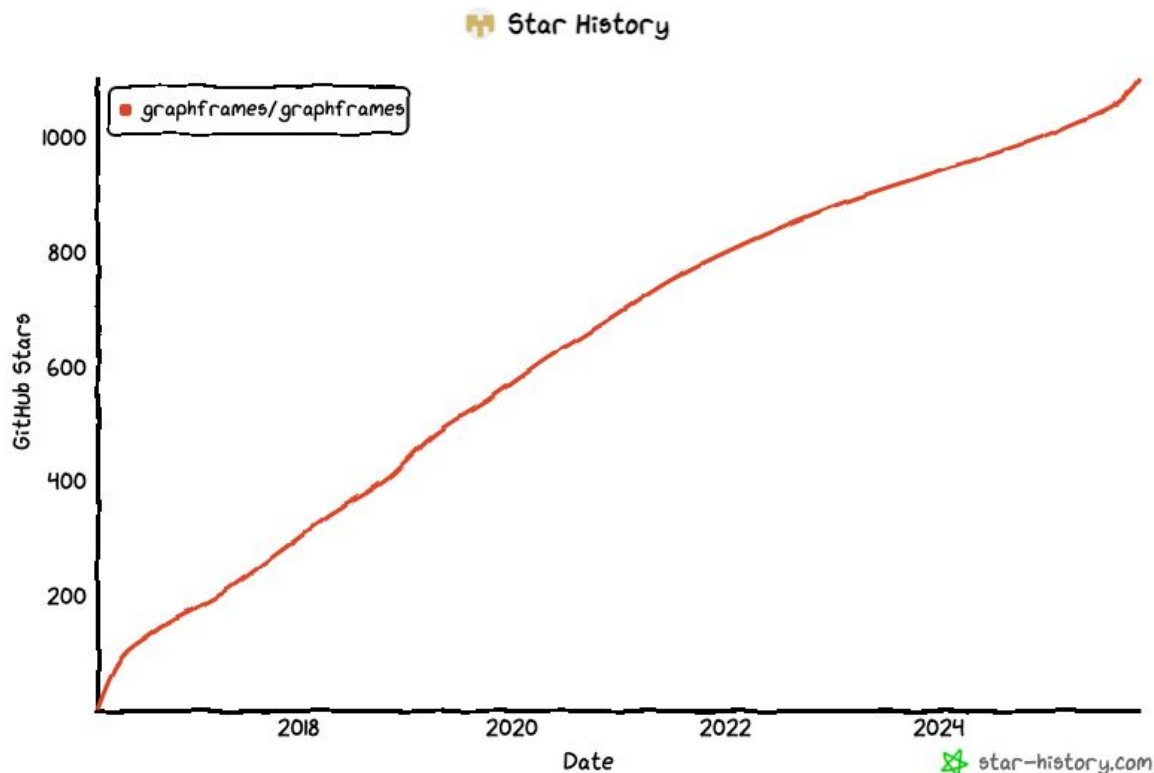
[4]: <https://whimsy.apache.org/board/minutes/Spark.html#2024-11-20>





GraphFrames

GraphFrames is collaborative effort among UC Berkeley, MIT, and Databricks



"After a six year gap in additions, GraphFrames is back with Spark Connect support!"

MVP

AML

Tasks

1. Link identities

User → profile → recipients

2. Connect participants

User 1 → User 2

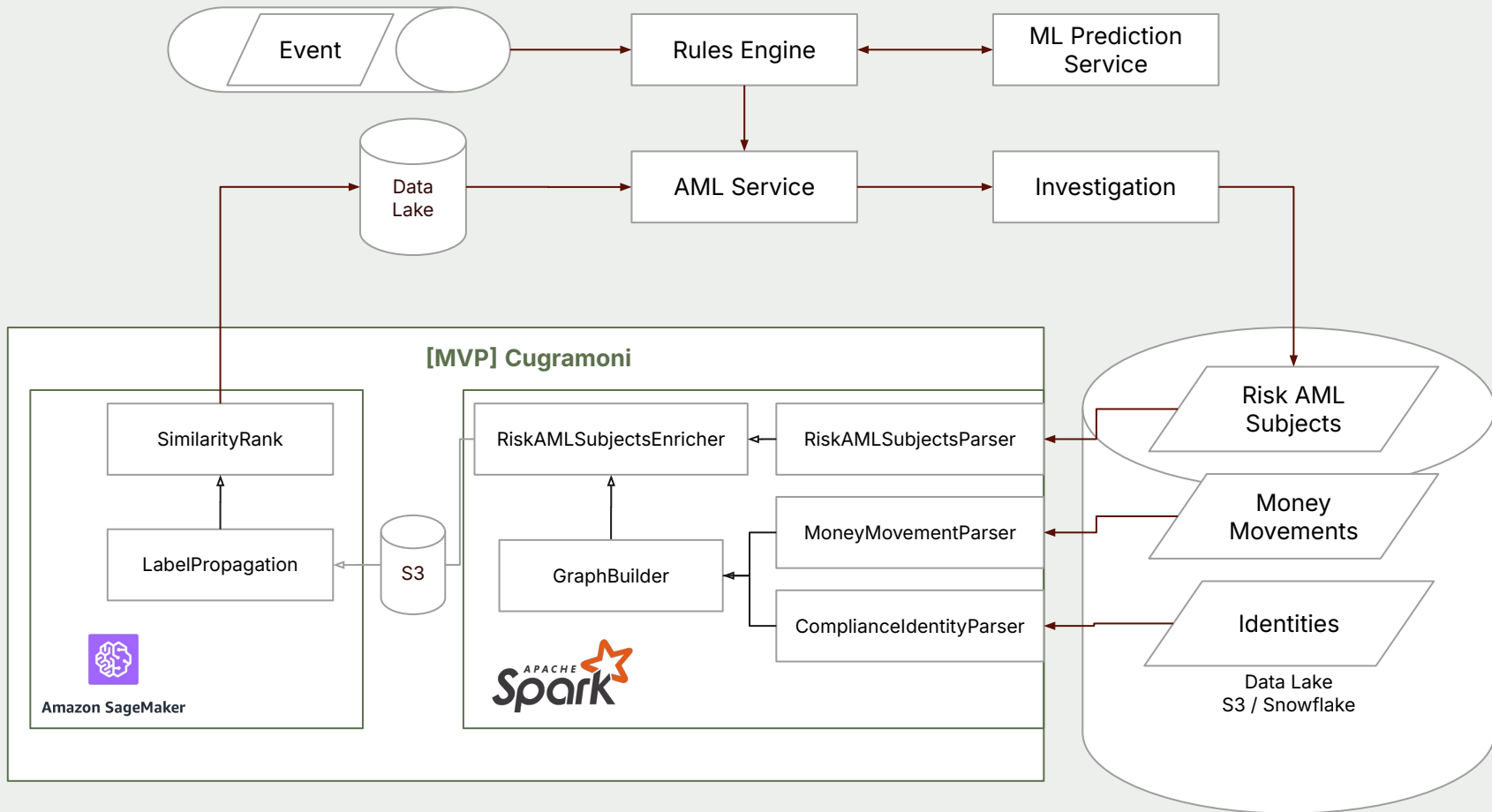
3. Add connected participants to AML Case

Reported User 1 → New User 2

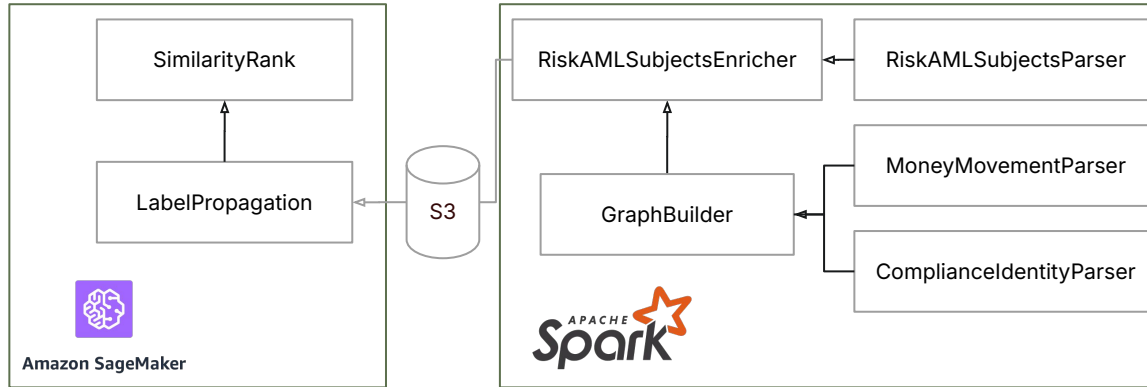
4. Filter top-k participants

5. Create new cases for investigation

ARCHITECTURE



[MVP] Cugramoni



DEMO

Why don't you use Neo4j for that?

Misses the point

Tools



Solutions

NEXT STEPS

LIVE CODE

github.com/ottobricks/demo-graphs-spark

INTERESTED?

JOB RESULTS

39 results



Mohammed Topiwalla

Lead Data
Scientist - Fraud
Prevention ☆



London



Analytics



Luca Traverso

Director of Data
Science -
Servicing ☆



London



Analytics



Pavlo Tertychnyi

Lead Data
Scientist - Spend ☆



Tallinn



Analytics





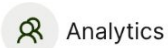
Rain Vagel



Senior Data Scientist- Risk



Tallinn



Analytics



Alexandr Severinov



Senior Data Science Manager - Financial...



London



Analytics



Lead Data Scientist - Treasury Liquidity



London



Analytics





Pere Morell

Marketing Science Manager

 London

 Analytics



Senior Data Scientist - Marketing

 London

 Analytics



Lead Marketing Scientist

 London

 Analytics





It's a me!

Senior ML Engineer

Senior Data Scientist

Tallinn

Analytics

→

